Applied Statistics With R
Exercises: Logit and Probit Models

John Fox

WU-Wien May/June 2006

It would be much too time-consuming to do all of these problems. A suggestion: Pick one of the binary logistic-regression problems and, if you wish, one of the other problems.

1 Binary Logistic Regression

1.1 Mroz's Data on Women's Labour-Force Participation

The data in the data frame `Mroz` in the `car` package, originally employed by Mroz (1987), were used by Long (1997) to illustrate the method of logistic regression. The variables in the dataset are described in the following table, adapted from Long (and using Long's variable names). Most of the variables are self-explanatory. The variable `lwg` is the log of the wife's actual wage if she is working outside the home. For women not working outside the home, Mroz proceeded as follows: He regressed the log-wages of the working women on the other variables, and used the resulting regression equation to predict the wages of those not working working outside the home.

Variable	*Description*
`lfp`	1 if the wife is in the paid labour force, 0 otherwise
`k5`	number of children ages 5 and younger
`k618`	number of children ages 6 to 18
`age`	Wife's age in years
`wc`	1 if wife attended college, 0 otherwise
`hc`	1 if husband attended college, 0 otherwise
`lwg`	log of wife's estimated wage rate
`inc`	family income excluding wife's wages, $1000s

Following Long, perform a logistic regression of `lfp` on the other variables. Briefly (i.e., in a paragraph) summarize the results of this regression. Offer *two* concrete interpretations of the coefficient of `inc` in the logistic regression.

1.2 Powers and Xie's Data on High-School Graduation

Employing a sample of 1643 men between the ages of 20 and 24 from the U. S. National Longitudinal Survey of Youth, Powers and Xie (2000) investigate the relationship between high-school graduation and parents' education, race, family income, number of siblings, family structure, and a test of academic ability. The data set, in the file `Powers.txt` on the course web site, contains the following variables (using Powers and Xie's variable names):

`hsgrad`	Whether the respondent was graduated from high school by 1985 (`Yes` or `No`).
`nonwhite`	Whether the respondent is black or Hispanic (`Yes` or `No`).
`mhs`	Whether the respondent's mother is a high-school graduate (`Yes` or `No`).
`fhs`	Whether the respondent's father is a high-school graduate (`Yes` or `No`).
`income`	Family income in 1979 (in $1000s) adjusted for family size.
`asvab`	Standardized score on the Armed Services Vocational Aptitude Battery test.[1]
`nsibs`	Number of siblings.
`intact`	Whether the respondent lived with both biological parents at age 14 (`Yes` or `No`).

The data file also contains respondent ID numbers, which are not contiguous.

(a) Following Powers and Xie perform a logistic regression of `hsgrad` on the other variables in the data set. Compute a likelihood-ratio test of the omnibus null hypothesis that *none* of the explanatory variables influences high-school graduation. Then construct 95-percent confidence intervals for the coefficients of the seven explanatory variables. What conclusions can you draw from these results? Finally, offer *two* brief, but concrete, interpretations of *each* of the estimated coefficients of `income` and `intact`.

(b) The logistic regression in the previous problem assumes that the partial relationship between the log-odds of high-school graduation and number of siblings is linear. Test for nonlinearity by fitting a model that treats `nsibs` as a factor, performing an appropriate likelihood-ratio test. In the course of working this problem, you should discover two errors in the data. Deal with the errors in a reasonable manner. Does the result of the test change?

2 Polytomous Logistic Regression

2.1 General Social Survey Data on Attitudes Toward Working Mothers

This data set is analyzed by Long (1997). The response variable has four ordered categories, Strongly Disagree, Disagree, Agree, and Strongly Agree, in relation to the statement, "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work." The explanatory variables are the year of the survey (1977 or 1989), the gender of the respondent, the race of the respondent (white or nonwhite), the respondent's age, and the prestige of the respondent's occupation (a quantitative variable). The data are in `WorkingMoms.txt` on the course web site.

(a) Following Long, perform an ordered (proportional odds) logistic regression of attitude toward working mothers on the other variables. What conclusions do you draw?

(b) Assess whether the proportional-odds assumption appears to hold for this regression. Fit a multinomial logit model to the data, and compare and contrast the results with those from the proportional odds model.

(c) Consider that possibility that gender interacts with the other explanatory variables in influencing the response variable. What do you find?

[1] Apparently scores on this test were standardized to a mean of 0 and standard deviation of 1 for the NLSY sample as a whole; as you can verify, the mean and standard deviation in this subsample differ somewhat from 0 and 1, respectively.

3 Logit Models for Contingency Tables

3.1 Harris, Luginbuhl, and Fishbein's Data on Invasions of Personal Space

Harris, Luginbuhl, and Fishbein (1978) conducted a social-psychological experiment that examined reactions to invasions of personal space. The research took place in a field setting provided by a public escalator. The primary results of the study were presented in the form of a contingency table (below). Three of the variables in the table were design or explanatory variables: (1) density of people on the escalator, rated as either high or low; (2) the sex of the subject; and (3) the sex of the intruder. The fourth variable was a dichotomous response variable: whether or not the subject reacted in some manner to the intrusion.

The authors analyzed the data by separately examining the two-way (partial) tables relating density to response within combinations of categories of the other three variables. Because there is a statistically significant relationship between density and response in only one of the four partial tables, the authors concluded that "males in the present study were more likely to react to a personal space invasion under low-density conditions than high-density conditions, but only when the intruder was another male. Density had no effect on responses by female subjects" (Harris, Luginbuhl, and Fishbein, 1978: 352-353). The implication is that there is a three-way interaction among the explanatory variables in determining response.

(1) Density	*(2) Sex of Subject*	*(3) Sex of Intruder*	*(4) Response* Yes	No
Low	Male	Male	18	1
		Female	15	8
	Female	Male	17	5
		Female	12	7
High	Male	Male	13	6
		Female	16	4
	Female	Male	10	9
		Female	14	6

(a) Calculate the response-variable odds within combinations of explanatory-variable categories. Compute and graph the log-odds, commenting on the results.

(b) Construct an analysis-of-deviance table, testing the various interactions and main effects of the explanatory variables on response. Do these tests square with the descriptive findings in part (a)?

(c) On the basis of the tests in part (b), fit a final logit model that incorporates only those effects shown to be important (and, of course, effects marginal to them). Using the parameter estimates for the model, calculate and graph the fitted logits.

(d) Test for independence between density and response separately in each of the four partial tables. (You may either fit a logit model to each table, or perform a traditional Pearson chi-square test of independence.) Do you obtain the results reported by Harris, Luginbuhl, and Fishbein.

(e) Do the results of your logit analysis support the authors' conclusions [replicated in part (d)]? Which analysis to you prefer? Why?

3.2 The Berkeley Graduate-School Admissions Data

The "Berkeley graduate-school admissions data," in the following table, first reported by Bickel et al. (1975), have become a staple of textbooks on categorical data analysis because they so clearly illustrate the distinction between marginal and partial relationships. The table gives the numbers of students admitted to graduate school and those not admitted by department and gender of the applicant. The six departments at the University of California at Berkeley are not identified, and simply are denoted A, B, ..., F.

		Admission	
Department	*Gender*	Yes	No
A	Male	512	313
	Female	89	19
B	Male	353	207
	Female	17	8
C	Male	120	205
	Female	202	391
D	Male	138	279
	Female	131	244
E	Male	53	138
	Female	94	299
F	Male	22	351
	Female	24	317

Here is a two-way table for the same data showing the marginal relationship between gender and admission (i.e., collapsed over department):

	Admission	
Gender	Yes	No
Male	1198	1493
Female	557	1278

6. Working from the collapsed two-way table, calculate the log-odds (logits) of admission for males and for females. Then, working from the full three-way table, calculate the log-odds of admission for males and for females in each department. Graph the log-odds, putting departments on the horizontal axis and drawing separate profiles for males and females. Focusing on male-female differences, compare the log-odds calculated by departments with the log-odds collapsed over departments. What do you conclude?

7. Fit a logit model to the data in the three-way table, treating admission as the response variable; include main effects for gender and department, and the interactions between gender and department. Using likelihood-ratio tests (i.e., constructing an analysis of deviance table), test for each of these effects. What do you conclude?